Phenotypic overlap in the contribution of individual genes to CNV pathogenicity revealed by cross-species computational analysis of single-gene mutations in humans, mice and zebrafish

Doelken, Sandra C; Köhler, Sebastian; Mungall, Christopher J; Gkoutos, Georgios V; Ruef, Barbara J; Smith, Cynthia; Smedley, Damian; Bauer, Sebastian; Klopocki, Eva; Schofield, Paul N

Published in:
Disease Models & Mechanisms (DMM)

DOI:
10.1242/dmm.010322

Publication date:
2013

Citation for published version (APA):
Doelken, S. C., Köhler, S., Mungall, C. J., Gkoutos, G. V., Ruef, B. J., Smith, C., Smedley, D., Bauer, S., Klopocki, E., & Schofield, P. N. (2013). Phenotypic overlap in the contribution of individual genes to CNV pathogenicity revealed by cross-species computational analysis of single-gene mutations in humans, mice and zebrafish. *Disease Models & Mechanisms (DMM)*, *6*(2), 358-372. https://doi.org/10.1242/dmm.010322

Disease Models & Mechanisms 6, 358-372 (2013) doi:10.1242/dmm.010322

Phenotypic overlap in the contribution of individual genes to CNV pathogenicity revealed by cross-species computational analysis of single-gene mutations in humans, mice and zebrafish

Sandra C. Doelken[1,2,*], Sebastian Köhler[1,2,*], Christopher J. Mungall[3,*], Georgios V. Gkoutos[4], Barbara J. Ruef[5], Cynthia Smith[6], Damian Smedley[7], Sebastian Bauer[1], Eva Klopocki[1,2], Paul N. Schofield[6,8], Monte Westerfield[5], Peter N. Robinson[1,2,9,‡] and Suzanna E. Lewis[3,‡]

SUMMARY

Numerous disease syndromes are associated with regions of copy number variation (CNV) in the human genome and, in most cases, the pathogenicity of the CNV is thought to be related to altered dosage of the genes contained within the affected segment. However, establishing the contribution of individual genes to the overall pathogenicity of CNV syndromes is difficult and often relies on the identification of potential candidates through manual searches of the literature and online resources. We describe here the development of a computational framework to comprehensively search phenotypic information from model organisms and single-gene human hereditary disorders, and thus speed the interpretation of the complex phenotypes of CNV disorders. There are currently more than 5000 human genes about which nothing is known phenotypically but for which detailed phenotypic information for the mouse and/or zebrafish orthologs is available. Here, we present an ontology-based approach to identify similarities between human disease manifestations and the mutational phenotypes in characterized model organism genes; this approach can therefore be used even in cases where there is little or no information about the function of the human genes. We applied this algorithm to detect candidate genes for 27 recurrent CNV disorders and identified 802 gene-phenotype associations, approximately half of which involved genes that were previously reported to be associated with individual phenotypic features and half of which were novel candidates. A total of 431 associations were made solely on the basis of model organism phenotype data. Additionally, we observed a striking, statistically significant tendency for individual disease phenotypes to be associated with multiple genes located within a single CNV region, a phenomenon that we denote as pheno-clustering. Many of the clusters also display statistically significant similarities in protein function or vicinity within the protein-protein interaction network. Our results provide a basis for understanding previously un-interpretable genotype-phenotype correlations in pathogenic CNVs and for mobilizing the large amount of model organism phenotype data to provide insights into human genetic disorders.

INTRODUCTION

Genomic disorders make up a family of genetic diseases that are characterized by large genomic rearrangements, including deletions, duplications and inversions of specific genomic segments. Many such rearrangements result in the loss or gain of specific genomic segments and thus are referred to as copy number variants (CNV). These regions can contain multiple genes. The phenotypic abnormalities seen in diseases associated with CNVs are thought to be related to altered gene dosage effects in most cases (Branzei and Foiani, 2007). In assessing the medical relevance of a CNV for a patient with a range of observed phenotypic abnormalities, it is essential to ascertain whether the CNV is causative for the disease and/or is merely incidental. If the CNV is, in fact, the cause of the disease, it is then important to know which of the genes located within the CNV are associated with which of the phenotypic features. In this study we focus on the latter challenge.

At present, information on Mendelian disorders that are associated with about 2000 human genes is available from sources such as OMIM (Online Mendelian Inheritance in Man) (Hamosh et al., 2005). However, substantially more information is available from model organisms such as the mouse and the zebrafish (Schofield et al., 2012). Furthermore, it has previously been shown that model organism phenotype data can be used for the analysis of human CNV disorders. For instance, Webber and co-workers investigated CNVs associated with mental retardation by linking the genes in these CNVs with phenotypes found in mouse gene-knockout models and showed that pathogenic mental-retardation-associated CNVs are significantly

[1]Institute for Medical and Human Genetics, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany
[2]Max-Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany
[3]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[4]Department of Computer Science, University of Aberystwyth, Old College, King Street, Aberystwyth, SY23 2AX, UK
[5]ZFIN, University of Oregon, Eugene, OR 97403-5291, USA
[6]The Jackson Laboratory, Bar Harbor, ME 04609, USA
[7]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
[8]Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, CB2 3EG, UK
[9]Berlin Center for Regenerative Therapies (BCRT), Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany
*These authors contributed equally to this work
‡Authors for correspondence (peter.robinson@charite.de; SELewis@lbl.gov)

Received 7 June 2012; Accepted 15 October 2012

TRANSLATIONAL IMPACT

Clinical issue

More than 60 disease syndromes, covering a wide range of systems, have been associated with copy number variation (CNV) in the human genome. With the advent of whole genome sequencing, many more CNVs are being found in patients with previously unreported phenotypes. With currently available approaches, it is difficult to determine whether these CNVs cause the disease phenotype, or whether dosage effects of certain genes in the segment are responsible for specific aspects of the disease. Moreover, there are more than 5000 human genes about which nothing is known phenotypically, but for which detailed phenotypic information about their orthologs in model organisms is available. This study introduces a novel computational method for prioritizing candidate human disease genes using model organism phenotype data, and is applicable on a genome-wide scale.

Results

For 27 recurrent CNV disorders, the authors identify 802 gene-phenotype associations. Approximately half of these associations were previously reported and half were novel candidate associations. 431 associations were made solely on the basis of model organism phenotype data. The authors also observed a striking and statistically significant tendency for individual disease phenotypes to be associated with multiple genes located within a single CNV region, a phenomenon that they denote as 'pheno-clustering'. Many of the members in a given cluster display statistically significant similarity in protein function, or vicinity within the protein-protein interaction network.

Implications and future directions

This study represents proof-of-principle for using phenotype data from model organisms to augment human clinical data to establish candidate disease genes. This method can, in principle, be extended to help with prioritizing candidates in GWAS and other association studies. In summary, this method represents an important new computational application for model organism phenotype data, and is expected to be widely applicable for interpreting individual genotype and precision phenotype data emerging from personalized medicine.

enriched with genes whose mouse orthologs, when disrupted, result in a nervous system phenotype (Boulding and Webber, 2012; Hehir-Kwa et al., 2010; Webber et al., 2009).

The use of non-human models has proved to be one of the most powerful approaches to understanding human disease (Rosenthal and Brown, 2007; Schofield et al., 2010). The description of abnormal phenotypes to model organisms can inform our understanding of the pathogenicity of human mutations, help prioritize candidate genes identified from genome-wide association studies (GWAS) and other investigations, and help dissect complex disease syndromes (Boulding and Webber, 2012). For the mouse, we now have phenotypes for around 8500 genes, and 40,000 genotypes with phenotypic annotations, in the Mouse Genome Informatics database (Blake et al., 2011). All of these phenotypes are coded using the Mammalian Phenotype Ontology (MPO) (Smith and Eppig, 2009; Smith et al., 2005). For zebrafish, there are more than 60,000 phenotypic descriptions of many thousands of genotypes, encoded using Entity Quality syntax (Washington et al., 2009). The recently launched International Mouse Phenotyping Consortium and the Zebrafish Mutation Project, which aim to systematically phenotype knockout lines for every protein-coding gene in the mouse and zebrafish genomes (Bartsch et al., 2005; Brown and Moore, 2012), will provide the most comprehensive phenotypic descriptions of any higher organisms. Together with data already available in other model organism databases, this provides an increasingly rich resource that can be leveraged to understand the consequences of human mutation and functionally dissect the human genome. The barrier to computational use of these data has been the disparate and non-standardized way of describing human phenotypic data, which has traditionally relied on free text or terminologies designed for medical management, billing and epidemiology (Schofield et al., 2010). The advent of the Human Phenotype Ontology (HPO) (Robinson et al., 2008; Robinson and Mundlos, 2010) addresses these problems with human data and is increasingly being used by clinical geneticists and systems biologists; we are now in a position to address the cross-mining of phenotype data from humans and model organisms to enormous benefit. Cross-species ontological approaches that use computer reasoning over phenotype ontologies offer a promising new methodology to identify similarities between human disease manifestations and observations made in genetically modified model organisms (Hoehndorf et al., 2011; Mungall et al., 2010; Washington et al., 2009). Ontologies are knowledge representations that use controlled vocabularies designed to facilitate knowledge capture and computer reasoning (Robinson and Bauer, 2011). An ontology provides a computational representation of the concepts of a domain together with the semantic relations between them. The use of ontologies for phenotypic analysis is discussed in Schofield et al. and Gkoutos et al. (Schofield et al., 2011; Gkoutos et al., 2012).

In this study, we introduce a computational algorithm that takes advantage of computable definitions of human, mouse and zebrafish phenotypes to perform genome-wide interspecies phenotype comparisons to detect candidate genes in recurrent hereditary CNV disorders. We have computationally examined the relationships between phenotypes associated with recurrent CNV disorders and phenotypes associated with human and model organism single-gene diseases whose genes are located within the CNV intervals. We have identified a total of 802 candidate genes for individual phenotypic features, approximately half of which were not previously reported in the literature. We additionally found a striking tendency for individual phenotypic features in CNV disorders to be associated with two or more individual genes located within the CNV. In many cases, these genes share functions or are located in close proximity to one another within the protein interactome. Thus, our work provides a framework for the interpretation of CNV-associated phenotypes, suggesting that clustering of functionally related genes within CNVs might be an important factor related to the phenotypic abnormalities seen in affected individuals.

RESULTS

We have developed a computational framework to harness phenotypic information from model organisms and single-gene human hereditary disorders to gain insights into the genetic etiology of the complex phenotypes of CNV disorders (Fig. 1). The goal of our analysis was to identify genes located within CNVs that are most likely to be responsible for the individual phenotypic abnormalities of the disease (Table 1). All available phenotype data for humans, mice and zebrafish were integrated for this analysis, resulting in a total of 7546 phenotypically described gene families, including 5703 for which phenotypes were only available from the model organisms (Fig. 2). In total, 802 genes were identified for individual phenotypic features of 27 different recurrent CNV disorders, including 346 newly identified associations (Table 2; supplementary material Tables S1, S2). For the 27 CNV diseases

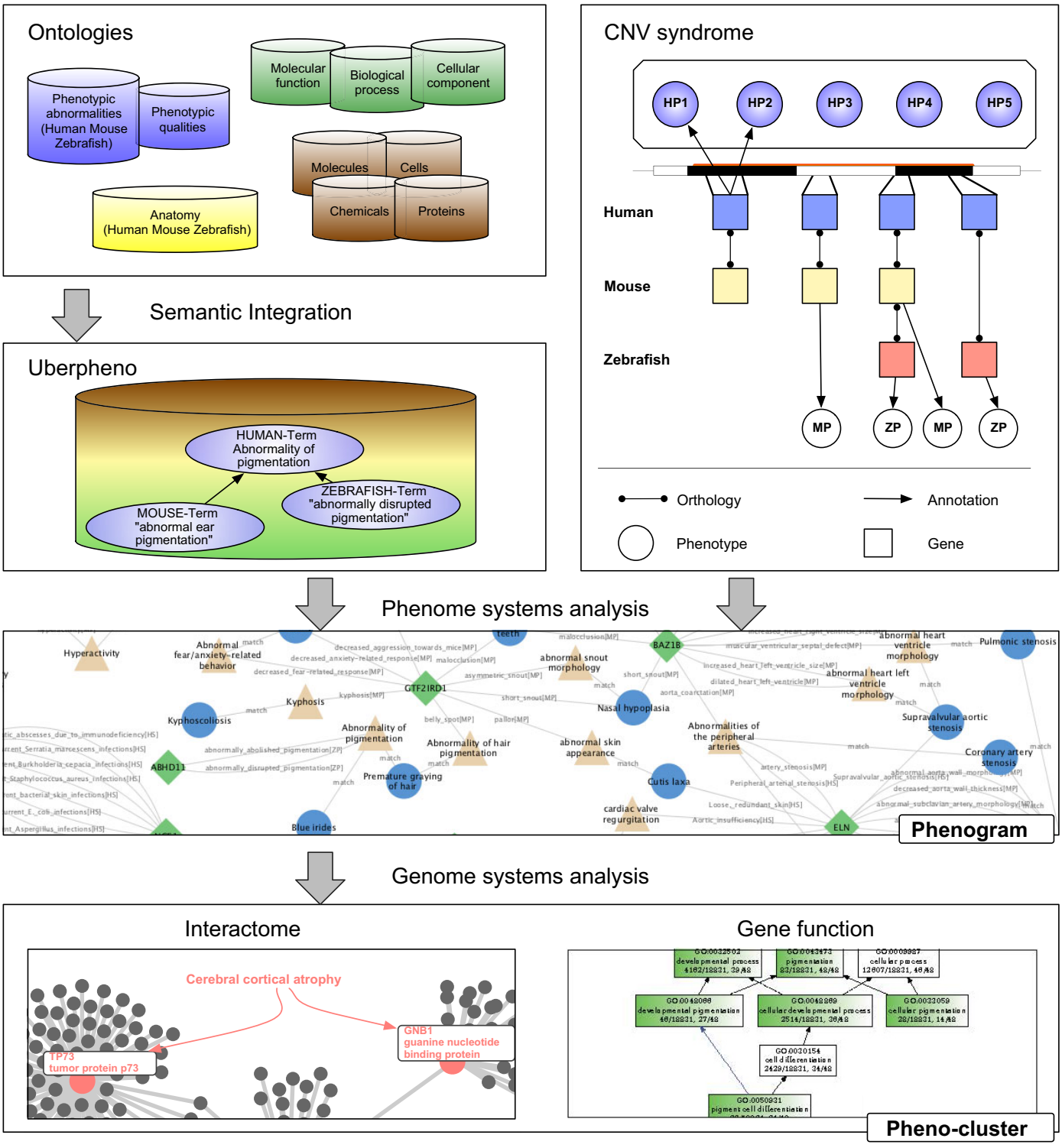

Fig. 1. A computational algorithm for genome-wide interspecies comparisons that detects candidate genes in CNV disorders. Information about phenotypic abnormalities in humans, mice and zebrafish together with the affected anatomical structures, functions, cells, chemicals and proteins is used to construct uberpheno, a phenome-wide network for cross-species comparison. To characterize the basis for phenotypic abnormalities seen in CNV disorders, phenotypic annotations for the CNVs, as well as for all single-gene disorders in humans and model organisms, is assembled together with links between orthologous genes. The figure shows an example CNV associated with five phenotypic features (HP1-HP5); the human genes located within the CNV, together with their orthologs in mouse and fish, are shown as squares. Arrows indicate a direct association derived from a database annotation and lines with circles connect orthologous genes. Phenome-wide analysis of these data is performed to characterize a 'phenogram', a network of genes and phenotypes for each CNV. Finally, the involved genes are analyzed with respect to functional similarity and vicinity in the protein interaction network.

Table 1. Summary of the 27 CNV disorders

Disease	ID[1]	Phenotype annotations *f*>15% (total)	Genes in CNV interval				*P*-value
			Total	Genes with phenotype information	Phenogram candidate genes	Previously reported phenogram candidate genes	
Xq28 (MECP2) duplication	D:45	14 (31)	23	12	6	2	<0.0002
NF1-microdeletion syndrome	MIM:613675	27 (32)	13	3	2	2	<0.0002
Leri-Weill dyschondrostosis	MIM:127300	13 (26)	1	1	1	1	0.0002
Familial adenomatous polyposis	MIM:175100	4 (19)	3	1	1	1	0.0002
WAGR 11p13 deletion syndrome	MIM:194072	9 (14)	5	2	2	2	0.0004
Pelizaeus-Merzbacher disease	MIM:312080	21 (26)	9	3	2	1	0.0004
Potocki-Shaffer syndrome	MIM:601224	23 (25)	15	9	4	2	0.0026
Split hand/foot malformation 1	MIM:183600	9 (11)	6	3	2	2	0.005
Sotos syndrome	MIM:117550	21 (38)	39	14	6	2	0.0122
Rubinstein-Taybi syndrome	MIM:180849	73 (112)	1	1	1	1	0.0138
Angelman syndrome	MIM:105830	25 (34)	50	9	7	3	0.0184
RCAD (renal cysts and diabetes)	MIM:137920	14 (23)	11	4	3	1	0.0216
Williams-Beuren syndrome	MIM:194050	68 (92)	34	13	11	4	0.0316
Wolf-Hirschhorn syndrome	MIM:194190	64 (81)	36	13	7	3	0.0478
Potocki-Lupski syndrome	MIM:610883	28 (32)	47	22	10	1	0.0628
9q subtelomeric deletion syndrome	MIM:610253	29 (38)	8	2	2	1	0.0662
Phelan-Mcdermid syndrome	MIM:606232	43 (54)	4	4	3	1	0.0728
Prader-Willi syndrome	MIM:176270	52 (66)	50	9	8	5	0.0788
17q21.3 microdeletion syndrome	MIM:610443	37 (51)	6	2	2	0	0.1094
Miller-Dieker syndrome	MIM:247200	41 (42)	37	21	15	9	0.1192
15q26 overgrowth syndrome	D:81	31 (37)	29	5	4	1	0.2028
1p36 microdeletion syndrome	MIM:607872	60 (86)	70	22	12	4	0.2762
Smith-Magenis syndrome	MIM:182290	40 (46)	47	22	13	4	0.2916
15q24 microdeletion syndrome	MIM:613406	56 (65)	36	15	8	1	0.2938
1q21.1 susceptibility locus (TAR)	MIM:274000	16 (44)	19	5	3	0	0.3178
Cri du Chat syndrome	MIM:123450	48 (68)	42	21	11	0	0.3374
3q29 microduplication syndrome	MIM:611936	14 (22)	22	6	2	1	0.5156

[1]OMIM (MIM); DECIPHER (D).

Included are the total number of phenotypic annotations with a frequency threshold of 15% (f>15%), as well as (in parentheses) the total number of annotations for that CNV disorder; the numbers of genes per CNV interval: total number of genes in the interval; number of genes with phenotype information (this is the number of genes that were included in the analysis); the number of phenogram candidate genes (i.e. those genes with phenotypic features in single-gene diseases in humans or model organisms that are similar to the features of the CNV disorders) identified in our study; the number of phenogram candidate genes that have been previously reported in the literature for the CNV disorder, as well as the empirical *P*-values for 10,000 randomizations (rows are sorted by *P*-value). The individual phenogram candidate genes as well as literature references for the previously reported gene-phenotype correlations are shown in supplementary material Table S1; for a detailed list of gene-phenotype associations see also supplementary material Table S2.

investigated in this work, a total of 468 of their respective phenotypic features could be explained on the basis of phenotypes associated with 802 single-gene mutations in humans, mice or zebrafish. The analysis consisted of four primary tasks: construction of logical definitions of pre-composed phenotype ontology terms, cross-species integration, calculation of information content of individual phenotype terms, and statistical comparison of phenotypes. For the frequent cases in which multiple individual genes provide a potential explanation for a phenotypic feature, the relationship of the genes to one another in the protein interactome and Gene Ontology functional space was examined (Fig. 1; see Materials and Methods, and supplementary material Table S3, which contains further details of the methods used).

The integration is possible because the computable definitions make use of the more atomic and species-agnostic elements from which each of the individual phenotypic classes is formed. For example, the HPO term 'increased bone mineral density' is composed of the Quality (PATO) term 'increased density' and the human anatomy term 'bone organ', which itself is a subclass of the Uberon (cross-species anatomy ontology) class 'bone'. A variety of such atomic ontologies covering biological processes, small molecules, cell types and anatomical structures are used to construct the logical (i.e. computable) definitions. In the second phase, our algorithm traverses the phenotype ontologies to integrate information from humans, mice and zebrafish semantically into a single composite ontology: 'uberpheno'. Thus, we combine information about the phenotype of

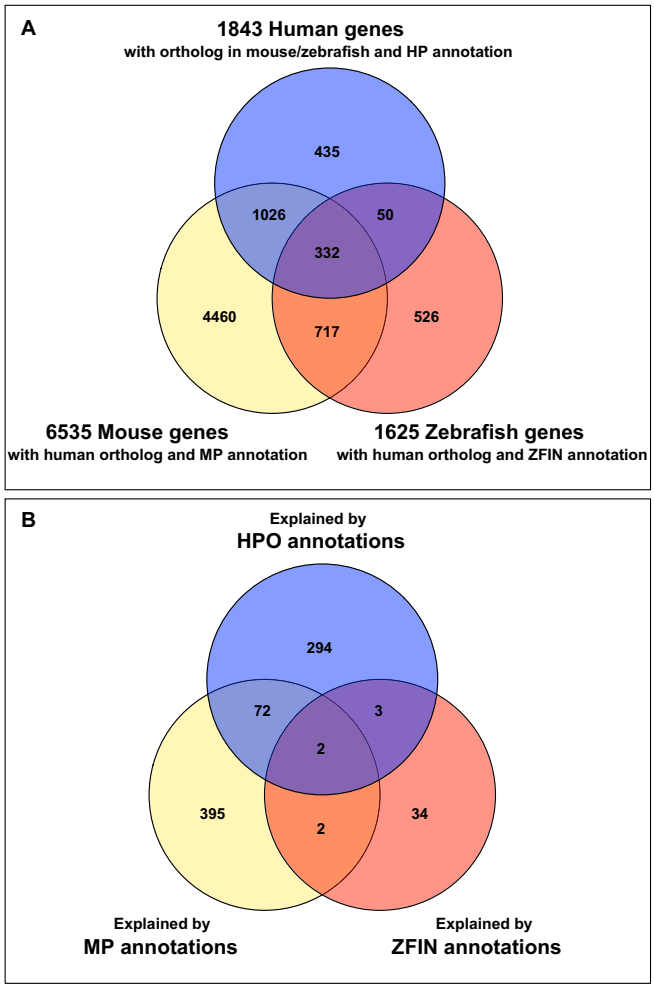

Fig. 2. Phenogram data and sources of explanations. (A) Venn diagram displaying the total numbers of human, mouse and zebrafish orthologous genes with phenotypic information. There are 5703 genes with phenotype data in mouse or zebrafish for which no human phenotype data are available. (B) Sources of the explanations of the 802 predicted phenotypic features of the 27 CNV disorders examined in this work. 431 of the 802 predictions were made only on the basis of model organism data (54%).

a human CNV disorder, the phenotypes (if known) of every human gene contained within the various possible CNVs, and the phenotypes of each of the gene orthologs in mouse and zebrafish. In the third step, the information content (IC) (Shannon, 1948) is calculated for each of the phenotypes in the composite ontology to provide a measure of how informative those annotations are. The IC is based on the number of genes annotated to the term in humans, mice and zebrafish. Very specific terms are associated only with very few genes and therefore have a high IC; nonspecific phenotypes associated with many genes have a low IC (Resnik, 1995). In the comparison step, each of the phenotypic descriptions for each of the genes is searched for similarity to each of the CNV phenotypes. Every phenotype scoring above a similarity threshold is selected to be part of the resulting 'phenogram' (described below; see also 'Quantification of phenogram score' in the Materials and Methods section), and each phenogram is scored according to its IC power sum. An empirical *P*-value is calculated for a phenogram by performing the entire analysis 5000 times with the same CNV phenotypes, but using randomized sets of genes. Our results are highly statistically significant; compared with the analyzed 27 CNV diseases, a total of ~480 of their respective phenotypic features could be explained, whereas in the randomized experiments, the average number of phenotypic features explained was ~250 (Fig. 3A).

For a number of the 27 CNVs analyzed, a critical gene is already known or highly suspected. Our results are highly consistent with these previous findings, supporting our computational methodology and interpretation, and we could additionally identify many previously unknown gene-phenotype connections (Table 3). The label 'major gene' in Table 3 refers to CNV disorders for which one gene is known to be the major player in the disease because intragenic mutations lead to a very similar phenotype in which most of the major phenotypic abnormalities characterizing the disease are present. It is common for individuals with the CNV disorder to present with more severe and/or additional features compared with individuals with intragenic single-gene mutations. Phenotypic variability of course also occurs within both cohorts of individuals – those with point mutations and those with the CNV disorder – which can make it difficult to pinpoint the systematic phenotypic differences between single-gene and CNV disorders. The underlying difficulty is also apparent in the current use of nomenclature: patients with the CNV disorder and patients with a point mutation in the 'main' disease-causing gene are defined as having 'Sotos syndrome' or 'neurofibromatosis' or 'Smith-Magenis syndrome', for example. Certainly, an individual with a microdeletion at 5q35 can be said to have Sotos syndrome but, even though most major features of the syndrome are present in cases of microdeletion and intragenic mutation, more or less obvious differences in phenotypic severity and additional abnormalities might be observed. In these cases, our method indicates where such additional effects might be present and offers information on additional genes and their potential contribution to certain phenotypic features of the disease.

Inclusion of data from different species into analyses such as ours is important because different model organisms can contribute distinct kinds of information. Our analysis results also support this concept, with interesting findings not just from mouse models but also from zebrafish: for the 15q26 overgrowth syndrome, micrognathia might be associated with haploinsufficiency of the gene *CHSY1*, an association drawn as a result of the zebrafish phenotype 'abnormally decreased size mandibular arch skeleton', whereas sensorineural hearing impairment might be a result of haploinsufficiency of *IGF1R*, which causes 'abnormally absent inner ear hair cell' in zebrafish models. Future projects will aim to include additional model organism data. The need for phenotype comparisons with further species has also been made evident only recently by the discovery of the major disease-causing gene for the 17q21.3 microdeletion syndrome, *KANSL1*. For this gene, no information was available from human, mouse or zebrafish, but it has been shown in 2010 by Lone et al. that *Drosophila* mutants show defects in synaptic vesicle biogenesis and trafficking, resulting in reduced learning ability (Lone et al., 2010).

Phenograms: prediction and visualization of genotype-phenotype associations

A phenogram of a CNV represents the network of genes and related phenotypes that have been associated with the genes in a particular

Table 2. Details of type and origins of results

	Novel gene-phenotype associations				Previously reported gene-phenotype associations			
Disease	**Total**	**HS**	**MP**	**ZP**	**Total**	**HS**	**MP**	**ZP**
Xq28 (MECP2) duplication	0	0	0	0	28	26	5	0
NF1-microdeletion syndrome	1	0	0	1	22	16	13	0
Leri-Weill dyschondrostosis	0	0	0	0	9	9	0	0
Familial adenomatous polyposis	0	0	0	0	4	4	4	0
WAGR 11p13 deletion syndrome	0	0	0	0	9	9	4	1
Pelizaeus-Merzbacher disease	1	0	0	1	10	8	6	0
Potocki-Shaffer syndrome	6	1	5	1	11	7	5	0
Split hand/foot malformation 1	0	0	0	0	15	0	15	0
Sotos syndrome	2	0	2	0	27	27	0	2
Rubinstein-Taybi syndrome	4	0	4	0	19	0	19	0
Angelman syndrome	9	2	7	0	38	32	11	0
RCAD (renal cysts and diabetes)	8	0	8	0	7	3	2	2
Williams-Beuren syndrome	23	6	15	2	16	4	14	0
Wolf-Hirschhorn syndrome	47	4	43	1	40	29	22	1
Potocki-Lupski syndrome	10	0	10	0	29	28	6	0
9q subtelomeric deletion syndrome	4	0	4	0	2	0	2	0
Phelan-Mcdermid syndrome	6	0	5	1	14	6	9	1
Prader-Willi syndrome	46	15	33	0	17	7	13	0
17q21.3 microdeletion syndrome	14	9	6	0	1	0	1	0
Miller-Dieker syndrome	18	0	15	3	28	9	22	0
15q26 overgrowth syndrome	14	0	8	6	7	7	1	1
1p36 microdeletion syndrome	67	26	35	8	28	13	13	2
Smith-Magenis syndrome	19	0	19	0	32	21	12	0
15q24 microdeletion syndrome	15	1	11	4	13	12	2	0
1q21.1 susceptibility locus (TAR)	7	0	7	0	0	0	0	0
Cri du Chat syndrome	23	1	21	1	29	28	1	1
3q29 microduplication syndrome	2	0	0	2	1	0	1	0
Σ = 802 Gene-phenotype associations	**346**	65	258	31	**456**	305	203	11

Included are the numbers of gene-phenotype associations, both novel and previously reported in the literature, as well as the corresponding numbers for the origins of the associations (HS, human; MP, mouse; ZP, zebrafish). In total, 802 gene-phenotype associations were made, corresponding to 346 novel and 456 previously reported associations. Note that the sum of the numbers of the origins for the predictions can be higher than the total numbers for novel or previously reported associations, because for some gene-phenotype associations there is evidence from more than one model organism, or from model organisms and other human diseases. All numbers correspond to single phenotype-to-gene associations – because one gene can be a candidate for more than one phenotypic abnormality, the numbers of these associations are higher than the total candidate gene numbers. For a list of candidate genes for each CNV disorder, see supplementary material Table S1; for a detailed list of all the gene-phenotype associations, see supplementary material Table S2.

CNV interval. All phenotype matches above a threshold, calculated based on the phenotype IC of the closest match (Resnik, 1995), are used to form a phenogram. For example, Rubinstein-Taybi syndrome is thought to result from haploinsufficiency of a single gene and, unsurprisingly, 19 of the phenotypic features of Rubinstein-Taybi syndrome were all assigned to CREBBP (Fig. 4). A more representative phenogram is observed in the analysis for Williams syndrome: from the original 68 frequent phenotypes and 13 phenotypically described genes in the CNV, the analysis generated a profile of 32 phenotypic abnormalities connected to 11 candidate genes through 39 associations. These included 16 of 23 previously reported associations (Pober, 2010) and 23 novel associations (Fig. 5). Note that, for Williams syndrome, we identified many phenotypic features that were associated with more than one candidate gene, i.e. multiple different genes all produce a similar phenotype.

In addition, the results indicate that dosage effects of other genes in an interval can present as modifying factors. For instance, the 9q subtelomeric deletion syndrome, clinically characterized by mental retardation, childhood hypotonia and facial dysmorphism, has been thought to result from haploinsufficiency of *EHMT1*, because point mutations in this gene cause a similar phenotype (Verhoeven et al., 2010). Our analysis identifies *CACNA1B* as a potential additional contributor to sleep disturbances and behavioral problems found in affected individuals. Similarly, point mutations in *HNF1B* are a known cause of RCAD (renal cysts and diabetes), but, in cases of microdeletions, haploinsufficiency of *ACACA*, which plays a role in glucose homeostasis, might

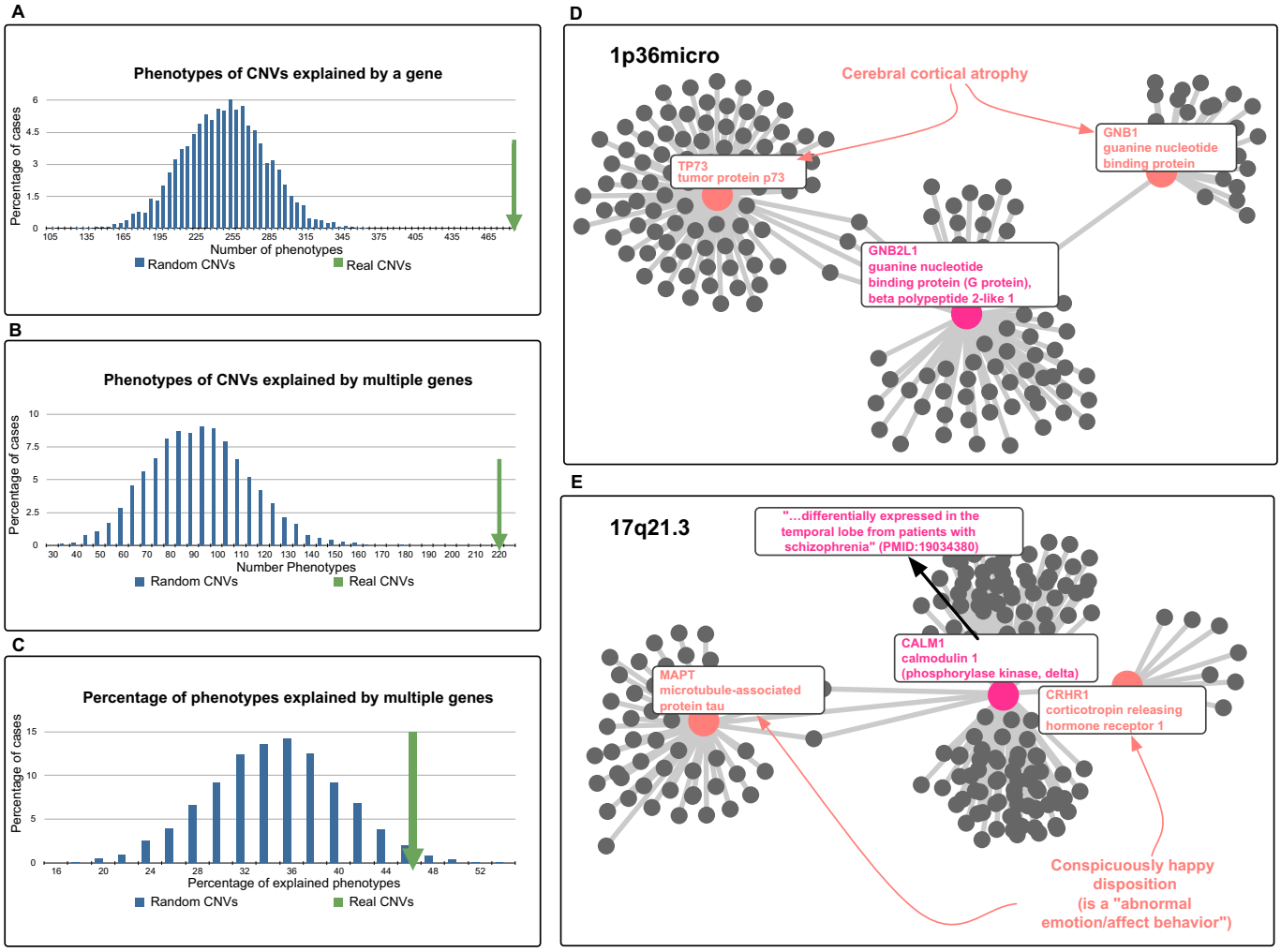

Fig. 3. Phenome/genome systems analysis. (A-C) Histograms illustrating the distribution of phenotypes with phenogram matches for the 27 CNVs investigated in this study (green arrow) versus randomly chosen CNVs (blue bars). (A) Number of phenotypes explained by one gene for randomly generated versus real CNVs. (B) Number of phenotypes explained by multiple genes (pheno-clusters) for randomly generated versus real CNVs. (C) Percentage of phenotypes explained by multiple genes as a percentage of all matching candidate genes for randomly generated versus real CNVs. All results in A-C are statistically significant. Results in C support the conclusion that pheno-clusters are not a characteristic of randomly chosen chromosomal segments (P=0.02). (D,E) Two pheno-clusters in which the genes in question are closely linked in the protein interaction network are shown. Approximately 20% of patients with 1p36 microdeletions present with 'cerebral cortical atrophy' (HP:0002120) (Battaglia et al., 2008). Mouse models of *TP73* are annotated to 'abnormal cerebral cortex morphology' (MP:0000788) and mouse models of *GNB1* to 'thin cerebral cortex' (MP:0006254); both associations were identified by our analysis. TP73 and GNB1 are closely linked in the protein interaction network through the gene *GNB2LI* (D). Approximately 50% of patients with 17q21.3 microdeletions are reported to present with a 'conspicuously happy disposition' (HP:0100024), which is an 'abnormal emotion/affect behavior' (HP:0100851). Mouse models of *MAPT* and *CRHR1* both present with a 'decreased anxiety-related response' (MP:0001364), which is also an 'abnormal emotion/affect behavior' (MP:0002572). Both genes are linked through the gene calmodulin 1 (*CALM1*) in the protein interaction network. Interestingly, Martins-de-Souza et al. (Martins-de-Souza et al., 2009) found that CALM1 is differentially expressed in the temporal lobe of patients with schizophrenia. Our results indicate that a combined dosage effect of MAPT and CRHR1, via modulating effects on CALM1, might explain some of the behavioral abnormalities observed in patients with 17q21.3 microdeletions.

contribute to the diabetes phenotype, and haploinsufficiency of *LHX1* might modify susceptibility to renal abnormalities.

Phenograms for all 27 CNV disorders investigated in this work can be downloaded from http://compbio.charite.de/tl_files/groupmembers/koehler/.

Pheno-clusters: composite effects of genes in CNV regions

A striking observation of our analysis was that numerous CNV phenotypes could be clearly associated with multiple genes in the interval, each of which in isolation has been shown to result in a similar phenotypic abnormality. Such pheno-clusters – physical clusters of genes associated with particular shared phenotypes in the genome – might be causative for a larger subset of the phenotypes observed in CNV disorders. Even genes that do not show dosage effects in isolation might cause phenotypic abnormalities if one or more additional pathway members are simultaneously deleted. Such an effect has been observed for the *SHFM1* locus, where the genes *DLX5* and *DLX6* are known to cause split-hand/split-foot malformation (SHFM). Existing mouse models exhibit the SHFM phenotype only if both genes are knocked out

Table 3. Comparison of results with previous findings.

Disease	Current knowledge of CNV etiology	Analysis results
Xq28 (MECP2) duplication	Major gene: *MECP2* (Ariani et al., 2004)	*MECP2* was recovered; additional candidates
NF1-microdeletion syndrome	Major gene: *NF1* (Venturin et al., 2004)	*NF1* was recovered; additional candidate
Leri-Weill dyschondrostosis	Major gene: *SHOX* (Rappold et al., 2002)	*SHOX* was recovered
Familial adenomatous polyposis	Major gene: *APC* (Groden et al., 1991)	*APC* was recovered
WAGR 11p13 deletion syndrome	Multigenic: *PAX6*; *WT1* (Fischbach et al., 2005) and additional candidates	*PAX6* and *WT1* were recovered; additional candidates
Pelizaeus-Merzbacher disease	Major gene: *PLP1* (Boespflug-Tanguy et al., 1994)	*PLP1* was recovered; additional candidate
Potocki-Shaffer syndrome	Multigenic	Suspected candidates recovered; novel candidates identified
Split hand/foot malformation 1	Digenic: *DLX5; DLX6* (Robledo et al., 2002)	*DLX5* and *DLX6* were recovered
Sotos syndrome	Major gene: *NSD1* (Kurotaki et al., 2002)	*NSD1* was recovered; additional candidates
Rubinstein-Taybi syndrome	Major gene: *CREBBP* (Hennekam, 2006; Tanaka et al., 1997)	*CREBBP* was recovered
Angelman syndrome	Major gene: *UBE3A* (Clayton-Smith and Laan, 2003; Jiang et al., 1999)	*UBE3A* was recovered; additional candidates
RCAD (renal cysts and diabetes)	Major gene: *HNF1B* (Bellanné-Chantelot et al., 2004; Bingham et al., 2001)	*HNF1B* was recovered; additional candidates
Williams-Beuren syndrome	Multigenic	Suspected candidates recovered; novel candidates identified
Wolf-Hirschhorn syndrome	Multigenic	Suspected candidates recovered; novel candidates identified
Potocki-Lupski syndrome	Multigenic, or duplication of *RAI1* ? (Potocki et al., 2007)	*RAI1* was recovered; additional candidates
9q subtelomeric deletion syndrome	Major gene: *EHMT1* (Kleefstra et al., 2005)	*EHMT1* was recovered; additional candidates
Phelan-Mcdermid syndrome	Major gene: *SHANK3* (Durand et al., 2007)	*SHANK3* was recovered; additional candidates
Prader-Willi syndrome	Multigenic	Suspected candidates recovered; novel candidates identified
17q21.3 microdeletion syndrome	Major gene: *KANSL1* (Zollino et al., 2012)	Previously suspected candidates recovered; novel candidates identified; *KANSL1* not recovered – no phenotype information from any organism!
Miller-Dieker syndrome	Multigenic	Suspected candidates recovered; novel candidates identified
15q26 overgrowth syndrome	Multigenic	Suspected candidates recovered; novel candidates identified
1p36 microdeletion syndrome	Multigenic	Suspected candidates recovered; novel candidates identified
Smith-Magenis syndrome	Major gene: *RAI1* (Girirajan et al., 2005; Slager et al., 2003)	*RAI1* was recovered; other suspected and novel candidates identified
15q24 microdeletion syndrome	Multigenic	Suspected candidates recovered; novel candidates identified
1q21.1 susceptibility locus (TAR)	Major gene: *RBM8A* (Albers et al., 2012)	Novel candidates identified; *RBM8A* not recovered –no phenotype information from any organism!
Cri du Chat syndrome	Multigenic	Novel candidates identified
3q29 microduplication syndrome	Multigenic	Suspected candidates recovered; novel candidate identified

The column 'Current knowledge of CNV etiology' indicates whether the literature on the CNV considers the etiology to be related to the effects of dosa ge alteration of a single major gene or to the combined effects of multiple genes. The term 'major gene' is used to refer to CNV disorders in which one gene is known to be the major factor determining the disease phenotype because, in most cases, point mutations in that gene lead to a very similar phenotype. For most of these CNV disorders, this one gene is known to be 'causative' for the main features, but most patients with the CNV disorder present with more severe and/or additional features compared with patients with an intragenic single-gene mutation. The column 'Analysis results' summarizes the results of the current study, indicating whether the known candidate genes were recovered and whether our algorithm identified new candidate genes for the phenotypic abnormalities of the CNV.

(Merlo et al., 2002; Robledo et al., 2002). Similarly, studies in murine models of Williams syndrome indicate that deletion of the *Eln* gene combined with the presence (non-deletion) of *Ncf1* contributes to the observed murine hypertension (Adams and Schmaier, 2012). The interaction between different genes affected by a CNV is potentially a very important determinant of clinical severity.

We investigated whether such phenotypic summation effects due to pheno-clusters occur more often than would be expected by

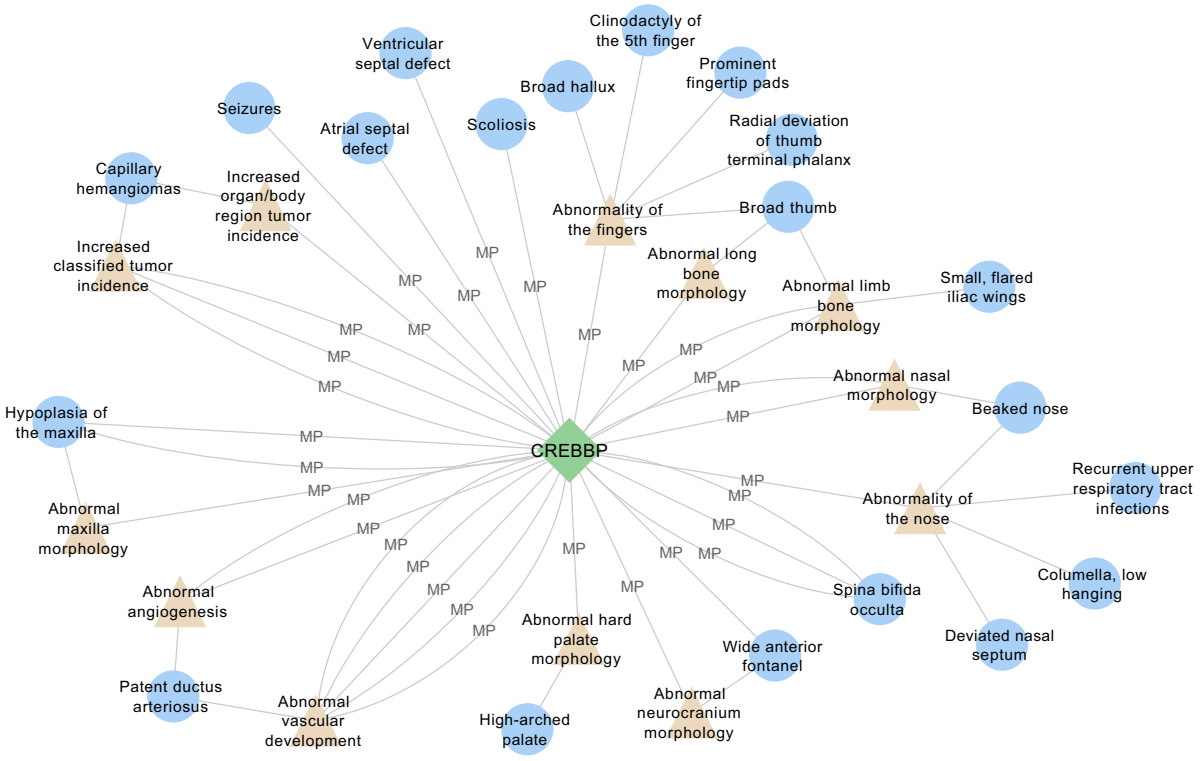

Fig. 4. Phenogram showing 19 phenotypic features of Rubinstein-Taybi syndrome assigned to a single gene, CREBBP, based solely on mouse phenotype data. See Bartsch et al. (Bartsch et al., 2005). CNV features are depicted as blue circles, genes as green diamonds and uberpheno terms as beige triangles. Relationships (shown as edges; connecting lines) between features and genes are labeled with the source of the inferred association: human (HS), mouse (MP) or zebrafish (ZP). Multiple edges (connections) between a gene and a CNV feature mean that multiple human, mouse or zebrafish phenotype descriptions (as labeled HS, MP or ZP) of single-gene disorders are similar to the phenotype observed in the CNV – for the precise terms of all of these phenotype matches see supplementary material Table S2. In contrast to Williams syndrome (see Fig. 5), all of the phenotypic features are attributable to a single gene. The phenogram was visualized using Cytoscape (Smoot et al., 2011). The phenograms of all 27 CNV disorders studied here can be downloaded from http://compbio.charite.de/tl_files/groupmembers/koehler/.

chance. The total number of pheno-clusters in our analysis was more than twice that of the randomized data (Fig. 3B). Because this could in part be a consequence of the lower overall number of genes and phenotypes identified in the randomized data, we examined the percentage of phenotypes in the randomized data explained by multiple genes. Even here, the percentage of pheno-clusters was significantly greater for the real CNVs than expected by random chance (P=0.02; Fig. 3C). In all, pheno-clusters were predicted for 220 phenotypes, corresponding to 135 gene clusters (in some cases, the same genes were associated with more than one phenotype). Thus, from the total of 802 gene-phenotype predictions, 220 (~25%) were explained by multiple genes, corresponding to pheno-clusters. It is known that the chromosomal location of genes can be related to their function. For instance, genes located adjacent to gene deserts very often function as transcriptional regulators (Ovcharenko et al., 2005). We therefore asked whether the functions of the genes identified in the pheno-clusters tend to functionally cluster as well. We analyzed the functional similarity of genes within each of the 135 pheno-clusters based on Gene Ontology criteria (see Materials and Methods). Overall, 49 of the pheno-clusters demonstrated a statistically significant intracluster similarity. We also used random walk analysis to investigate whether the gene products of the genes in the 135 pheno-clusters were in closer proximity in the protein interactome than expected by chance (Köhler et al., 2008). In total, 27 of the pheno-clusters showed a statistically significant vicinity score (see Materials and Methods), and 62 of the pheno-clusters (46%) were validated by both methods. Fig. 3D,E shows examples of statistically significant protein-protein interaction (PPI) network results. Although these examples are not based on experimental evidence, we note that explanations of CNV phenotypes currently given in the literature are almost exclusively based on 'guilt-by-association' from manual literature searches. In contrast, our results are based on a comprehensive phenome-wide search across data from humans and two model organisms. Our results also indicate the importance of collecting detailed phenotype-genotype information on patients with different forms of CNV diseases (i.e. due to point mutations in a single gene, and due to microdeletions), because this information could be relevant to their clinical management. Similar conclusions apply to CNV disorders characterized by variable intervals. For example, in Phelan McDermid syndrome, *SHANK3* was initially thought to be responsible for most of the phenotype, because it was included in the minimal critical region (Bonaglia et al., 2001). Our analysis suggests that *MAPK81P2* might contribute to the behavioral and autistic features of the disease. Not all individuals with this disease have deletions encompassing *SHANK3*. Thus, either *SHANK3* or *MAPK81P2* might cause behavioral problems, and combined haploinsufficiency of both genes might increase the likelihood of autism. Currently, publicly available data

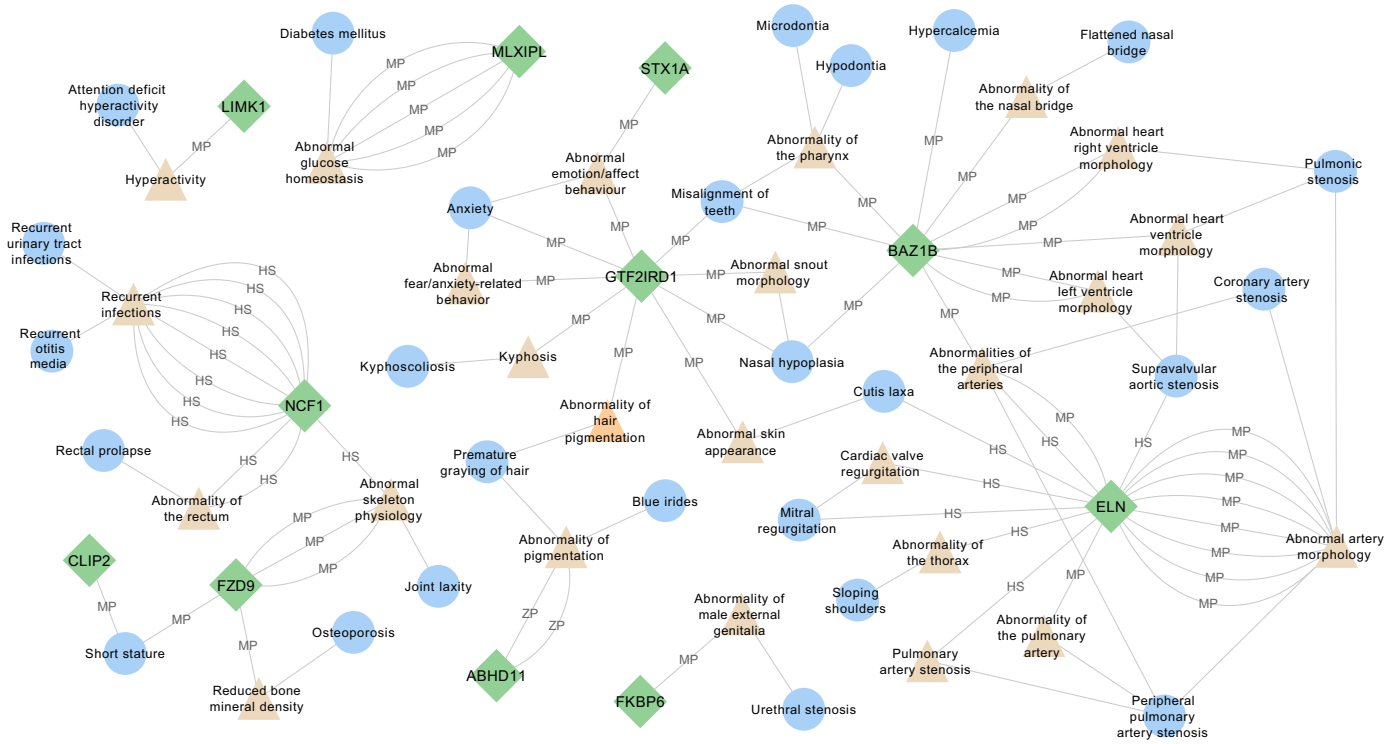

Fig. 5. Phenogram reveals 11 genes associated with 26 phenotypic abnormalities in Williams syndrome. See Fig. 4 for an explanation of the symbols. Williams syndrome is known as one of the classical contiguous gene syndromes, where the phenotypic features are thought to be caused by haploinsufficiency of a number of genes in the deleted interval. Some associations, for example the involvement of *ELN* in 'supravalvular aortic stenosis' or *BAZ1B* in 'hypercalcemia', have been previously reported. We identified many phenotypic features associated with more than one candidate gene. For example, *ELN* and *BAZ1B* are associated with phenotypic abnormalities of the cardiovascular system, *FZD9* and *CLIP2* are both associated with 'short stature', and *BAZ1B* and *GTF2IRD* with 'nasal hypoplasia' and 'misalignment of teeth'. *STX1A* has been previously associated with 'abnormal glucose tolerance' in humans; our results show that *MLXIPL* is also associated with 'abnormal glucose homeostasis', an association that has not been reported for Williams syndrome to date. Our method further detects behavioral similarities between species and predicts effects of *GTF2IRD1* on increased 'anxiety', and *LIMK1* on 'hyperactivity' and 'attention deficits', according to mouse models. Pigmentation abnormalities ('blue irides' and 'premature graying of hair') could be a side effect of the *GTF2IRD1* deletion, according to zebrafish models, and 'recurrent infections' seen in patients are associated with the deletion of *NCF1*.

do not include sufficient information about the phenotype and the extent of the deletions to draw this conclusion. Our analysis thus motivates further specific genotype-phenotype studies for CNV disorders such as Phelan McDermid syndrome that are characterized by variable interval sizes.

DISCUSSION

In clinical genetics, it is often difficult to decide whether a quantitative variation in the genome is related to the observed phenotype, and predicting consequences of haploinsufficiency is challenging (Huang et al., 2010). To understand the functional impact of a given CNV region, not only does the general issue of pathogenicity need to be answered, but also the question of which of the genes included in the CNV region are associated with which phenotypic abnormalities present in the patient. Such information is invaluable for clinical management. Patients with overlapping but different sized deletions or duplications might present with different phenotypes that correlate with the affected genes. For patient follow-up and screening procedures, the information that one patient might, for example, have a high cancer risk or a risk for developing diabetes or hypertension, whereas another patient does not, might have a huge impact on individual prognosis and treatment.

In this work, we developed a semantic algorithm for mapping model organism phenotype data to equivalent human phenotypic features. We used this algorithm to address the question of which genes in CNVs are most likely to be causally related to individual phenotypic features seen in the CNV based on the assumption that an abnormal dosage in a gene is likely to lead to similar phenotypic abnormalities as a loss- or gain-of-function mutation in the same gene. In this way, we are able to exploit the wealth of phenotypic information available for 5703 genes in model organisms for which phenotypes of mutations in the human orthologs are unknown (Fig. 2). For the 27 well-characterized CNV disorders analyzed in this work, we identified a total of 802 phenogram matches, i.e. genes in which a monogenic disease in humans or model organisms is associated with a phenotypic feature that is also seen in (or similar to) one of the features of the CNV disorder. In order to test the performance of our algorithm, we performed the identical analysis 5000 times on randomized data. On average, only 250 features were identified, and the maximum number of features found in any of the randomized runs was ~350 (Fig. 3A). We performed an extensive literature search for previously reported phenotype associations (supplementary material Table S2); comparison of the results of our algorithm revealed that we identified 457 previously

reported associations. Additionally, we found 346 novel phenotype associations that, to the best of our knowledge, have not been previously recognized in the medical literature. Our algorithms might additionally be valuable to incorporate model organism data into other areas in human genetics, such as the prioritization of variants found in exome sequencing projects.

Our work has several limitations. In Table 1, empirical probabilities (*P*-values) for the phenogram scores (S_{PG}) are given for each of the 27 CNV disorders. In total, 14 of the CNV disorders displayed statistically significant scores ($P<0.05$), including clinically distinct disorders such as WAGR and Sotos syndrome. There are several possible reasons for the lack of statistical significance of the remaining 13 disorders, which could relate either to the limitations of our computational approach, inadequate phenotypic annotations, or lack of knowledge about the genes located within the CNV. An important limitation of the approach as implemented in the current work is that it depends upon the granularity of the phenotype descriptions. More broadly used, nonspecific descriptions of abnormalities, such as autism or intellectual disability, are not flagged as 'statistically significant' because they are so frequently used. The calculated *P*-values are based on the IC of the phenotypic features, and the IC of intellectual disability is very low (IC=3.2) because so many genes (currently 392) are annotated to this term. Indeed, *P*-values of the phenogram scores reported in Table 1 correlate with the granularity of the phenotypic descriptions of the CNV disorders, shown as the average IC of the CNV phenotypes and the IC of the phenogram-matches; unsurprisingly, the *P*-values also correlate with the size of the intervals, measured with respect to the absolute numbers of genes and the numbers of genes with available phenotype information (see figure 3.2 in section 3.9 of supplementary material Table S3). Many recently characterized CNV disorders that have been delineated on the basis of array-CGH (comparative genomic hybridization) screening rather than clinical studies have substantially less-specific clinical pictures. Nonspecific clinical phenotypes and high phenotypic variability complicate diagnosis and could explain why diseases associated with microdeletions or duplications of 3q29 (Ballif et al., 2008) and microdeletions of 15q24 (Andrieux et al., 2009) do not score as well as more distinct CNV disorders. Although the current work concentrated on identifying statistically significant phenotypic matches, an implementation of our method as a clinical decision support system could be designed to show the best match or matches for both specific and less specific phenotypic abnormalities. We also note that the *P*-value as calculated in this work is not a measure of the probability that the CNV is the cause of the disease phenotype, which is the type of hypothesis testing that one would use in a diagnostic setting. Rather, the statistical hypothesis is a measure of whether the phenotypic abnormalities associated with the individual genes within the CNV match the phenotype of the CNV disorder better than one would expect by random chance, which is a conservative way of evaluating the results of semantic phenotype matching. It is to be expected that some degree of phenotypic similarity to CNVs with complex phenotypes exists at many other loci in the genome. For instance, hundreds of distinct CNVs could be associated with phenotypes such as autism (Levy et al., 2011), and indeed there is a high likelihood that a large deletion anywhere in the genome will be pathogenic and result in one or more abnormal phenotypic features (Vermeesch et al., 2007). Therefore, the method presented in this work would need to be extended to include other data, such as previous reports of comparable CNVs in databases such as DECIPHER (Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources) (Firth et al., 2009) and ISCA (Riggs et al., 2012), to be useful as a clinical differential diagnosis support tool.

It is difficult to provide any direct experimental proof in humans that altered dosage of a specific gene is responsible for a specific phenotypic abnormality in a CNV disorder, unless monogenic lesions also occur in isolation in other patients. Rather, candidate genes are proposed based on the similarity of their single-gene mutation phenotypes to the CNV phenotypes; for instance, haploinsufficiency for *ELN* was proposed as the cause of supravalvular aortic stenosis observed in individuals with Williams syndrome because point mutations in the *ELN* gene also give rise to this phenotype. A total of 456 of the 802 associations identified by computational analysis in our study have been previously proposed in the literature, thereby supporting our computational approach (supplementary material Table S2). To the best of our knowledge, 346 of the 802 associations offer novel candidate genes for individual phenotypic features. We examined associations from the literature that were not detected by our approach to determine the extent and possible reasons for such false-negatives. Some associations, such as *FZD9* and 'intellectual disability' for Williams syndrome (Pober, 2010), fell below the threshold for detection by our method because of their very low IC. Others, such as the association of genetic variants in *STX1A* and 'impaired glucose tolerance' (Pober, 2010; Romeo et al., 2008), were not detected by our method because they were based on human association studies and are so far not reported in OMIM or any of the information sources used in this study. Inclusion of these kinds of data, for example from resources such as the Genetic Association Database (GAD) (Becker et al., 2004) or GWAS Central (Thorisson et al., 2009), in candidate gene prediction algorithms such as ours will be addressed in future projects.

For neurological and neuropsychiatric phenotypes such as intellectual disability, seizures, schizophrenia, mood disorders and autism, genetic heterogeneity and variable expressivity and penetrance are well-known features. Cytogenetic imbalances are the most frequently identified cause of intellectual disability (Aradhya et al., 2007), and CNVs are increasingly being detected by array-CGH in individuals with neurological and neuropsychiatric phenotypes (Akil et al., 2010). For such phenotypes, dysregulation of relevant neural circuits might be caused by disruption of single genes, but combinatorial effects of variations in many genes affecting shared pathways have also been proposed (Shaikh et al., 2011). Similarly, clustering of functionally related genes has been proposed for bovine quantitative trait loci (Salih and Adelson, 2009). We identified clustering of functionally related genes within CNVs as a second important factor for pathogenicity of CNVs in the human genome, not only for neurological phenotypes, but also for various other phenotypic features such as genitourinary, skeletal and metabolic abnormalities. We found evidence that genes involved in pheno-clusters are often functionally related to one another and tend to be near one another in the PPI network (Fig. 3D,E).

There is abundant evidence now that there is functional clustering in all mammalian genomes. Presumably, the phenotypic clustering observed in our study is related to the clustering of functional

neighborhoods of genes across chromosomes, which is even partially conserved across species (Al-Shahrour et al., 2010). In some cases, clustering is associated with areas of strong linkage disequilibrium, suggesting that coinheritance of combinations of alleles of genes whose products interact or are associated with the same pathway or function might be the evolutionary driving force. Interestingly, functional clusters shared by different species do not always seem to consist of orthologs, suggesting that evolutionary pressure is exerted upon the cluster's function rather than the individual genes within it (Al-Shahrour et al., 2010; Michalak, 2008; Petkov et al., 2005). To date, there has been no explicit global analysis of the clustering of gene function, location, process, pathway or expression patterns involved in human CNVs, but the possibility of epistatic relationships between these genes would be predicted to be strong. There is, however, some evidence that certain functional gene classes are overrepresented in areas of the genome containing common CNVs (Conrad et al., 2010). Our data provides for the first time a breakdown of the 'phenotypic readout' from regions involved in CNVs and strongly suggests that they contain functionally clustered genes (Michalak, 2008; Petkov et al., 2005; Stranger et al., 2007). The results of our study shed new light on the pathobiology of human CNVs and provide evidence that the concept of clustering of phenotypically related genes plays an important role in genome pathology.

Another important aspect of our study is that 377 of the 629 genes analyzed did not have any human or model organism phenotype information. Thus, systematic genome-wide phenotyping efforts such as the International Mouse Phenotyping Consortium (Brown and Moore, 2012) and corresponding efforts in zebrafish (Kettleborough et al., 2011; Wang et al., 2007), such as the Zebrafish Mutation Project, have great potential to provide additional insights and candidates for genes involved in human disease. Algorithms such as ours that make use of phenotypic similarities between human and model organisms will facilitate the computational integration of information from these projects, harnessing these increasingly rich resources to help us understand the consequences of human mutation and functionally dissect the human genome. Our algorithms can be adapted to assist with interpretation and understanding of the diagnostic results from array-CGH analyses. Similar algorithms can be developed for interpreting next-generation sequencing data, thereby moving closer to the objective of a personalized genetic approach to medical care.

MATERIALS AND METHODS

Phenotype annotations using 'uberpheno', a cross-species phenotype ontology

We downloaded 17 ontologies from the Open Biological and Biomedical Ontologies (OBO) Foundry website (Smith et al., 2007) and constructed the logical definitions for HPO terms and MPO terms from these. The definitions can thereby serve to relate entities across the three species for common biological processes, small biological molecules and cell types. The anatomical terms used in the phenotype definitions, from the corresponding anatomical ontologies of the three species, were related to one another using the metazoan anatomy ontology Uberon (Mungall et al., 2012). Using these definitions, we created a single combined cross-species ontology called 'uberpheno' that represents phenotypes in mouse, human and zebrafish (Köhler et al., 2011). Full details of the construction of uberpheno are provided in supplementary material Table S3.

The phenotype ontologies do not themselves represent diseases, but rather describe individual phenotypic abnormalities. Any one disease may comprise one or more such abnormalities; therefore, each disease is represented computationally by an annotation to multiple phenotypic abnormalities. For this work, we compiled phenotype annotations from multiple sources. All available phenotypic information from humans was extracted from the HPO annotations, the majority of which are based on data from the OMIM knowledgebase (Amberger et al., 2009). Data on 6535 murine models were obtained from the Mouse Genome Informatics (MGI) database (Shaw, 2009), and detailed phenotypic annotations for 1625 zebrafish models were taken from ZFIN (Zebrafish Model Organism Database) (Bradford et al., 2011). The 27 recurrent CNV disorders examined in this work were manually curated using HPO terms to generate comprehensive sets of annotations. The main sources for the manual annotation were recent publications, GeneReviews (Pagon et al., 1993), OMIM (Amberger et al., 2009) and a standard reference work on dysmorphology in human genetics (Jones and Smith, 1997). The intervals and corresponding genes included in the intervals of the chosen CNVs were taken from DECIPHER (Firth et al., 2009). For this project, a conservative approach was chosen by including all genes in a maximal critical region as stated by DECIPHER. For some diseases, a gene that was not included by DECIPHER was added to the list for the corresponding CNV because of evidence from recent publications stating involvement of the gene. For detailed information on individual annotations including references for all annotated phenotypes, as well as the complete gene lists for the intervals of all 27 CNV disorders, see supplementary material Table S4.

Computational strategy for CNV analysis

The analysis of a CNV begins with the set of genes ($G_{CNV} = \{g_1, g_2, \ldots, g_n\}$) located within the CNV. For each of the genes ($g_i \in G_{CNV}$) there is a set (T_{g_i}) of associated phenotype terms from human single-gene disorders and from available mouse and zebrafish models. Similarly, T_{CNV} represents the set of phenotypes associated with the particular CNV disorder. Phenotype annotations for humans (HPO), mouse (MPO) and zebrafish (directly composed Entity and Quality annotations) are mapped to the corresponding terms in uberpheno. Phenotypic features that are only rarely associated with the CNV (i.e. less than 15% of affected persons show the feature) were removed from T_{CNV} before further analysis.

Information content of phenotype terms

The IC of a term t is defined as the negative logarithm of the frequency of annotations to the term (Resnik, 1995): $IC(t) = -\log p_t$, where p_t is the probability of annotations to term t of uberpheno among all annotated genes in humans, mice and zebrafish.

Common ancestors

We define $anc(\cdot)$ as a function that, for a given term or set of terms, returns the set of ancestral terms (i.e. inferred super-classes). Note, this function is reflexive, i.e. $t \in anc(t)$. The set of common ancestors of an uberpheno term T_{g_i} associated with gene g_i and the set of uberpheno terms associated with the CNV is defined as:

$$CA(t_{g_i}, T_{CNV}) = \left\{ t \middle| t \in \mathrm{anc}(T_{CNV}) \cap \mathrm{anc}(t_{g_i}) \right\}. \quad (1)$$

We define $t_{\max}(t_{g_i}, T_{CNV})$ to be a term with the highest information content from the set $CA(t_{g_i}, T_{CNV})$.

Phenograms

We define the phenogram of the CNV as a structure (G,P,D,E,λ) where G refers to the genes that have a matching phenotype above the IC threshold λ, P are the matching gene phenotypes, and D are the disorder phenotypes, with E the edges that connect them. G consists of all genes in G_{CNV} for which a single-gene phenotype t_g matches with a phenotype of the CNV with an information content above the threshold, i.e. $IC[t_{max}(t_g,T_{CNV})]\geq\lambda$ (these genes are shown as green squares in Figs 4 and 5). P consists of all phenotypes in $t_{max}(t_g,T_{CNV})$ (shown as beige triangles in Figs 4, 5). The genes are connected to one or more shared phenotype terms with the connections labeled according to the source organism (HP, MP or ZP). Finally, the triangles are connected to the phenotype terms of the CNV (T_{CNV}; shown as blue circles) that are explained by the matches. The phenograms were visualized using Cytoscape (Shannon et al., 2003; Smoot et al., 2011).

Quantification of phenogram score

For each gene, $g \in G_{CNV}$, a phenomatch score S_g is defined based on the information content of all matching terms with specificity above a certain similarity threshold λ to exclude relatively non-specific phenotypic features:

$$S_g(g,T_{CNV}) = \sum_{\substack{t_g \in T_g \\ IC(t_{\max}(t_g T_{CNV}))\geq\lambda}} [IC(t)]^k . \quad (2)$$

In our analysis, we selected λ to be 2.5, corresponding to a frequency for the feature of 786 among all 9580 analyzed genes. By choosing k>1, terms with a higher IC receive higher weighting. For this study, we selected k to be 5. The full phenogram score across all genes located in the CNVs can then be calculated as:

$$S_{PG}(G_{CNV},T_{CNV}) = \sum_{g\in G_{CNV}} S_g(g,T_{CNV}) . \quad (3)$$

We note that the HPO annotations for the 27 CNV disorders in this work were created by manual curation. Additionally, OMIM contains some entries that correspond to CNV diseases such as Rubinstein Taybi Deletion syndrome (MIM:610543). To avoid bias, these annotations were excluded when analyzing the corresponding CNV. Table 1 shows empirical P-values based on the S_{PG} scores of the 27 CNV disorders. To calculate the P-values, 5000 intervals containing the same number of genes as the original CNV were generated at random and S_{PG} was calculated. The P-value was estimated as the proportion of times in which the randomized interval scored at least as high as the original CNV.

Distribution of phenogram matches

Each of the phenotypic matches, i.e. genes (g) annotated to some term (t_g) whose similarity to a term in C exceeds the information content threshold λ, represents a potential 'explanation' of a phenotypic feature of the CNV. We reasoned that, although individual matches could be due to chance, the total number of above-threshold matches could provide a useful measure of the utility of our method. For the complete analysis, we included all genes from model organisms that have an ortholog in human as well as phenotype information. All human genes located within the CNV interval (G_{CNV}) were then analyzed as described above, and the number of terms in P were summed over all of the 27 CNV disorders under consideration. We calculated an empirical P-value for the distribution by keeping the set of CNVs and their phenotypic abnormalities fixed while comparing them to randomly chosen sets of genes (G_r) to replace the original set of genes (G). This was done by randomly selecting a human gene (g_r) and defining a random interval (G_r) surrounding g_r that contained the same number of human genes as the original CNV.

Pheno-clusters and functional relatedness

In our analysis, we identified groups of genes (G_p) located in the same CNV that are associated with the same phenotypic abnormality. We investigated the hypothesis that these 'pheno-clusters' of genes are not only related to the same phenotype but also share similarity based on other biological measures. Here, we calculated similarity based on Gene Ontology (GO) annotations of the genes in G_p and examined the vicinity of the genes to one another within PPI networks.

To compute the homogeneity of $G_p = \{g_1, g_2, \ldots, g_n\}$ based on GO, we compute the average pairwise similarity for all unique pairs of genes in G_p:

$$HOM_{GO}(G_p) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} sim_{GO}(g_i, g_j) . \quad (4)$$

For a pair of genes, we calculate the symmetric semantic similarity $[sim_{GO}(g_i, g_j)]$ as in equation 2 of Köhler et al. (Köhler et al., 2009). To determine a P-value for a given homogeneity score $[HOM_{GO}(G_p)]$, we set up the empirical score distribution by randomly generating 10,000 random gene groups (G_r) and computing $HOM_{GO}(G_r)$, then estimating the P-value as the fraction of cases in which $HOM_{GO}(G_r) \geq HOM_{GO}(G_p)$.

In order to test the hypothesis that genes in the same pheno-cluster also tend to cluster in the human PPI network, we analyzed a network containing 10,742 nodes, corresponding to human genes coding for proteins with known interactions, as described previously (Köhler et al., 2008). We constructed the column-normalized adjacency matrix **A** and then computed the random walk matrix **P** by $\boldsymbol{P} = [\boldsymbol{I} - (1-r)\boldsymbol{A}]^{-1} \times r$, where $\boldsymbol{I}$ is the identity matrix. Every entry (P_{ij}) represents the probability of a random walker starting at node i and being at node j after an infinite number of steps. In every step, the walker randomly visits adjacent nodes. Note that, with probability r, the walker is reset to the start node i. In our study we set r to 0.75.

For a group of genes ($G_p = \{g_1, g_2, \ldots, g_n\}$), we compute the average global network proximity $[GNP(G_p)]$ by:

$$GNP(G_p) = \frac{1}{|G_p|} \sum_{g_i \in G_p} p^i_\infty [g_i], \quad (5)$$

whereby p_∞ is calculated as $P \times p^i_o$. To set up the vector of start probabilities (p^i_o), the start probability of a network node k is defined as :

$$p^i_o[g_k] = \frac{1}{|G_p|-1}, \quad (6)$$

if $g_k \in \{G_p \backslash g_i\}$, and 0 otherwise. Thus, when analyzing a particular g_i, the random walker starts with equal probability from all nodes in G_p except g_i. Then, the random walk distance from all the start

nodes to g_i is computed and the average over all $g_i \in G_p$ is taken as the $GNP(G_p)$.

Similar to the GO analysis, we determine a *P*-value for a given score $GNP(G_p)$ by calculating the empirical score distribution. This was done by randomly generating 10,000 random gene groups (G_r) and computing $GNP(G_r)$. Afterwards, we estimate the *P*-value as the fraction of cases in which $GNP(G_r) \geq GNP(G_p)$.

COMPETING INTERESTS

The authors declare that they do not have any competing or financial interests.

AUTHOR CONTRIBUTIONS

P.N.R., S.E.L., P.N.S. and M.W. conceived, coordinated and supervised the study. S.K., C.J.M., S.C.D. and S.B. developed the computational methods. S.K. and S.C.D. performed the analysis and analyzed the data. S.C.D., S.K., C.J.M., G.V.G., B.J.R., C.S., D.S., E.K., P.N.R., P.N.S. and M.W. worked on ontology development and annotations. S.C.D., S.K., S.E.L., P.N.S. and P.N.R. wrote the paper.

FUNDING

This work was supported by the Director, Office of Science, Office of Basic Energy Sciences, of the US Department of Energy under Contract No. DE-AC02-05CH11231, and by grants of the Deutsche Forschungsgemeinschaft (DFG RO 2005/4-1), the Bundesministerium für Bildung und Forschung (BMBF project number 0313911), the MGD grant from the National Institutes of Health, HG000330, the ZFIN grant from the National Institutes of Health, U41-HG002659 and the PATO grant from the National Institutes of Health, R01-HG004838.

SUPPLEMENTARY MATERIAL

Supplementary material for this article is available at http://dmm.biologists.org/lookup/suppl/doi:10.1242/dmm.010322/-/DC1

REFERENCES

Adams, G. N. and Schmaier, A. H. (2012). The Williams-Beuren Syndrome-a window into genetic variants leading to the development of cardiovascular disease. *PLoS Genet.* **8**, e1002479.

Akil, H., Brenner, S., Kandel, E., Kendler, K. S., King, M. C., Scolnick, E., Watson, J. D. and Zoghbi, H. Y. (2010). Medicine. The future of psychiatric research: genomes and neural circuits. *Science* **327**, 1580-1581.

Al-Shahrour, F., Minguez, P., Marqués-Bonet, T., Gazave, E., Navarro, A. and Dopazo, J. (2010). Selection upon genome architecture: conservation of functional neighborhoods with changing genes. *PLOS Comput. Biol.* **6**, e1000953.

Albers, C. A., Paul, D. S., Schulze, H., Freson, K., Stephens, J. C., Smethurst, P. A., Jolley, J. D., Cvejic, A., Kostadima, M., Bertone, P. et al. (2012). Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nat. Genet.* **44**, 435-439.

Amberger, J., Bocchini, C. A., Scott, A. F. and Hamosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.* **37**, D793-D796.

Andrieux, J., Dubourg, C., Rio, M., Attie-Bitach, T., Delaby, E., Mathieu, M., Journel, H., Copin, H., Blondeel, E., Doco-Fenzy, M. et al. (2009). Genotype-phenotype correlation in four 15q24 deleted patients identified by array-CGH. *Am. J. Med. Genet. A.* **149A**, 2813-2819.

Aradhya, S., Manning, M. A., Splendore, A. and Cherry, A. M. (2007). Whole-genome array-CGH identifies novel contiguous gene deletions and duplications associated with developmental delay, mental retardation, and dysmorphic features. *Am. J. Med. Genet. A.* **143A**, 1431-1441.

Ariani, F., Mari, F., Pescucci, C., Longo, I., Bruttini, M., Meloni, I., Hayek, G., Rocchi, R., Zappella, M. and Renieri, A. (2004). Real-time quantitative PCR as a routine method for screening large rearrangements in Rett syndrome: Report of one case of MECP2 deletion and one case of MECP2 duplication. *Hum. Mutat.* **24**, 172-177.

Ballif, B. C., Theisen, A., Coppinger, J., Gowans, G. C., Hersh, J. H., Madan-Khetarpal, S., Schmidt, K. R., Tervo, R., Escobar, L. F., Friedrich, C. A. et al. (2008). Expanding the clinical phenotype of the 3q29 microdeletion syndrome and characterization of the reciprocal microduplication. *Mol. Cytogenet.* **1**, 8.

Bartsch, O., Schmidt, S., Richter, M., Morlot, S., Seemanová, E., Wiebe, G. and Rasi, S. (2005). DNA sequencing of CREBBP demonstrates mutations in 56% of patients with Rubinstein-Taybi syndrome (RSTS) and in another patient with incomplete RSTS. *Hum. Genet.* **117**, 485-493.

Battaglia, A., Hoyme, H. E., Dallapiccola, B., Zackai, E., Hudgins, L., McDonald-McGinn, D., Bahi-Buisson, N., Romano, C., Williams, C. A., Brailey, L. L. et al. (2008). Further delineation of deletion 1p36 syndrome in 60 patients: a recognizable phenotype and common cause of developmental delay and mental retardation. *Pediatrics* **121**, 404-410.

Becker, K. G., Barnes, K. C., Bright, T. J. and Wang, S. A. (2004). The genetic association database. *Nat. Genet.* **36**, 431-432.

Bellanné-Chantelot, C., Chauveau, D., Gautier, J. F., Dubois-Laforgue, D., Clauin, S., Beaufils, S., Wilhelm, J. M., Boitard, C., Noël, L. H., Velho, G. et al. (2004). Clinical spectrum associated with hepatocyte nuclear factor-1beta mutations. *Ann. Intern. Med.* **140**, 510-517.

Bingham, C., Bulman, M. P., Ellard, S., Allen, L. I., Lipkin, G. W., Hoff, W. G., Woolf, A. S., Rizzoni, G., Novelli, G., Nicholls, A. J. et al. (2001). Mutations in the hepatocyte nuclear factor-1beta gene are associated with familial hypoplastic glomerulocystic kidney disease. *Am. J. Hum. Genet.* **68**, 219-224.

Blake, J. A., Bult, C. J., Kadin, J. A., Richardson, J. E., Eppig, J. T. and the Mouse Genome Database Group (2011). The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.* **39**, D842-D848.

Boespflug-Tanguy, O., Mimault, C., Melki, J., Cavagna, A., Giraud, G., Pham Dinh, D., Dastugue, B. and Dautigny, A. (1994). Genetic homogeneity of Pelizaeus-Merzbacher disease: tight linkage to the proteolipoprotein locus in 16 affected families. PMD Clinical Group. *Am. J. Hum. Genet.* **55**, 461-467.

Bonaglia, M. C., Giorda, R., Borgatti, R., Felisari, G., Gagliardi, C., Selicorni, A. and Zuffardi, O. (2001). Disruption of the ProSAP2 gene in a t(12;22)(q24.1;q13.3) is associated with the 22q13.3 deletion syndrome. *Am. J. Hum. Genet.* **69**, 261-268.

Boulding, H. and Webber, C. (2012). Large-scale objective association of mouse phenotypes with human symptoms through structural variation identified in patients with developmental disorders. *Hum. Mutat.* **33**, 874-883.

Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Howe, D. G., Knight, J., Mani, P., Martin, R., Moxon, S. A. et al. (2011). ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res.* **39**, D822-D829.

Branzei, D. and Foiani, M. (2007). Template switching: from replication fork repair to genome rearrangements. *Cell* **131**, 1228-1230.

Brown, S. D. and Moore, M. W. (2012). Towards an encyclopaedia of mammalian gene function: the International Mouse Phenotyping Consortium. *Dis. Model. Mech.* **5**, 289-292.

Clayton-Smith, J. and Laan, L. (2003). Angelman syndrome: a review of the clinical and genetic aspects. *J. Med. Genet.* **40**, 87-95.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P. et al. (2010). Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-712.

Durand, C. M., Betancur, C., Boeckers, T. M., Bockmann, J., Chaste, P., Fauchereau, F., Nygren, G., Rastam, M., Gillberg, I. C., Anckarsäter, H. et al. (2007). Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. *Nat. Genet.* **39**, 25-27.

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R. M. and Carter, N. P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84**, 524-533.

Fischbach, B. V., Trout, K. L., Lewis, J., Luis, C. A. and Sika, M. (2005). WAGR syndrome: a clinical review of 54 cases. *Pediatrics* **116**, 984-988.

Girirajan, S., Elsas, L. J., 2nd, Devriendt, K. and Elsea, S. H. (2005). RAI1 variations in Smith-Magenis syndrome patients without 17p11.2 deletions. *J. Med. Genet.* **42**, 820-828.

Gkoutos, G. V., Schofield, P. N. and Hoehndorf, R. (2012). Computational tools for comparative phenomics: the role and promise of ontologies. *Mamm. Genome* **23**, 669-679.

Groden, J., Thliveris, A., Samowitz, W., Carlson, M., Gelbert, L., Albertsen, H., Joslyn, G., Stevens, J., Spirio, L., Robertson, M. et al. (1991). Identification and characterization of the familial adenomatous polyposis coli gene. *Cell* **66**, 589-600.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514-D517.

Hehir-Kwa, J. Y., Wieskamp, N., Webber, C., Pfundt, R., Brunner, H. G., Gilissen, C., de Vries, B. B., Ponting, C. P. and Veltman, J. A. (2010). Accurate distinction of pathogenic from benign CNVs in mental retardation. *PLoS Comput. Biol.* **6**, e1000752.

Hennekam, R. C. (2006). Rubinstein-Taybi syndrome. *Eur. J. Hum. Genet.* **14**, 981-985.

Hoehndorf, R., Schofield, P. N. and Gkoutos, G. V. (2011). PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res.* **39**, e119.

Huang, N., Lee, I., Marcotte, E. M. and Hurles, M. E. (2010). Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* **6**, e1001154.

Jiang, Y., Lev-Lehman, E., Bressler, J., Tsai, T. F. and Beaudet, A. L. (1999). Genetics of Angelman syndrome. *Am. J. Hum. Genet.* **65**, 1-6.

Jones, K. and Smith, D. (1997). *Smith's Recognizable Patterns of Human Malformation*. Philadelphia, PA: Saunders.

Kettleborough, R. N., Bruijn, E., Eeden, F., Cuppen, E. and Stemple, D. L. (2011). High-throughput target-selected gene inactivation in zebrafish. *Methods Cell Biol.* **104**, 121-127.

Kleefstra, T., Smidt, M., Banning, M. J., Oudakker, A. R., Van Esch, H., de Brouwer, A. P., Nillesen, W., Sistermans, E. A., Hamel, B. C., de Bruijn, D. et al. (2005).

Disruption of the gene Euchromatin Histone Methyl Transferase1 (Eu-HMTase1) is associated with the 9q34 subtelomeric deletion syndrome. *J. Med. Genet.* **42**, 299-306.

Köhler, S., Bauer, S., Horn, D. and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* **82**, 949-958.

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., Mundlos, C., Horn, D., Mundlos, S. and Robinson, P. N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* **85**, 457-464.

Köhler, S., Bauer, S., Mungall, C. J., Carletti, G., Smith, C. L., Schofield, P., Gkoutos, G. V. and Robinson, P. N. (2011). Improving ontologies by automatic reasoning and evaluation of logical definitions. *BMC Bioinformatics* **12**, 418.

Kurotaki, N., Imaizumi, K., Harada, N., Masuno, M., Kondoh, T., Nagai, T., Ohashi, H., Naritomi, K., Tsukahara, M., Makita, Y. et al. (2002). Haploinsufficiency of NSD1 causes Sotos syndrome. *Nat. Genet.* **30**, 365-366.

Levy, D., Ronemus, M., Yamrom, B., Lee, Y. H., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K. et al. (2011). Rare *de novo* and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886-897.

Lone, M., Kungl, T., Koper, A., Bottenberg, W., Kammerer, R., Klein, M., Sweeney, S. T., Auburn, R. P., O'Kane, C. J. and Prokop, A. (2010). The nuclear protein Waharan is required for endosomal-lysosomal trafficking in Drosophila. *J. Cell Sci.* **123**, 2369-2374.

Martins-de-Souza, D., Gattaz, W. F., Schmitt, A., Rewerts, C., Marangoni, S., Novello, J. C., Maccarrone, G., Turck, C. W. and Dias-Neto, E. (2009). Alterations in oligodendrocyte proteins, calcium homeostasis and new potential markers in schizophrenia anterior temporal lobe are revealed by shotgun proteome analysis. *J. Neural Transm.* **116**, 275-289.

Merlo, G. R., Paleari, L., Mantero, S., Genova, F., Beverdam, A., Palmisano, G. L., Barbieri, O. and Levi, G. (2002). Mouse model of split hand/foot malformation type I. *Genesis* **33**, 97-101.

Michalak, P. (2008). Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics* **91**, 243-248.

Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E. and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. *Genome Biol.* **11**, R2.

Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. and Haendel, M. A. (2012). Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* **13**, R5.

Ovcharenko, I., Loots, G. G., Nobrega, M. A., Hardison, R. C., Miller, W. and Stubbs, L. (2005). Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15**, 137-145.

Pagon, R. A., Bird, T. D., Dolan, C. R., Stephens, K. and Adam, M. P. (ed.) (1993). Seattle (WA): University of Washington, Seattle. *GeneReviews*.

Petkov, P. M., Graber, J. H., Churchill, G. A., DiPetrillo, K., King, B. L. and Paigen, K. (2005). Evidence of a large-scale functional organization of mammalian chromosomes. *PLoS Genet.* **1**, e33.

Pober, B. R. (2010). Williams-Beuren syndrome. *N. Engl. J. Med.* **362**, 239-252.

Potocki, L., Bi, W., Treadwell-Deering, D., Carvalho, C. M., Eifert, A., Friedman, E. M., Glaze, D., Krull, K., Lee, J. A., Lewis, R. A. et al. (2007). Characterization of Potocki-Lupski syndrome (dup(17)(p11.2p11.2)) and delineation of a dosage-sensitive critical interval that can convey an autism phenotype. *Am. J. Hum. Genet.* **80**, 633-649.

Rappold, G. A., Fukami, M., Niesler, B., Schiller, S., Zumkeller, W., Bettendorf, M., Heinrich, U., Vlachopapadoupoulou, E., Reinehr, T., Onigata, K. et al. (2002). Deletions of the homeobox gene SHOX (short stature homeobox) are an important cause of growth failure in children with short stature. *J. Clin. Endocrinol. Metab.* **87**, 1402-1406.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *International Joint Conference for Artificial Intelligence* (IJCAI-95), pp. 448-453.

Riggs, E. R., Jackson, L., Miller, D. T. and Van Vooren, S. (2012). Phenotypic information in genomic variant databases enhances clinical care and research: the International Standards for Cytogenomic Arrays Consortium experience. *Hum. Mutat.* **33**, 787-796.

Robinson, P. N. and Bauer, S. (2011). *Introduction to Bio-Ontologies*. 517pp. Boca Raton, FL: Taylor & Francis.

Robinson, P. N. and Mundlos, S. (2010). The human phenotype ontology. *Clin. Genet.* **77**, 525-534.

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D. and Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* **83**, 610-615.

Robledo, R. F., Rajan, L., Li, X. and Lufkin, T. (2002). The Dlx5 and Dlx6 homeobox genes are essential for craniofacial, axial, and appendicular skeletal development. *Genes Dev.* **16**, 1089-1101.

Romeo, S., Sentinelli, F., Cavallo, M. G., Leonetti, F., Fallarino, M., Mariotti, S. and Baroni, M. G. (2008). Search for genetic variants of the SYNTAXIN 1A (STX1A) gene: the -352 A>T variant in the STX1A promoter associates with impaired glucose metabolism in an Italian obese population. *Int. J. Obes. (Lond.)* **32**, 413-420.

Rosenthal, N. and Brown, S. (2007). The mouse ascending: perspectives for human-disease models. *Nat. Cell Biol.* **9**, 993-999.

Salih, H. and Adelson, D. L. (2009). QTL global meta-analysis: are trait determining genes clustered? *BMC Genomics* **10**, 184.

Schofield, P. N., Gkoutos, G. V., Gruenberger, M., Sundberg, J. P. and Hancock, J. M. (2010). Phenotype ontologies for mouse and man: bridging the semantic gap. *Dis. Model. Mech.* **3**, 281-289.

Schofield, P. N., Sundberg, J. P., Hoehndorf, R. and Gkoutos, G. V. (2011). New approaches to the representation and analysis of phenotype knowledge in human diseases and their animal models. *Brief. Funct. Genomics* **10**, 258-265.

Schofield, P. N., Hoehndorf, R. and Gkoutos, G. V. (2012). Mouse genetic and phenotypic resources for human genetics. *Hum. Mutat.* **33**, 826-836.

Shaikh, T. H., Haldeman-Englert, C., Geiger, E. A., Ponting, C. P. and Webber, C. (2011). Genes and biological processes commonly disrupted in rare and heterogeneous developmental delay syndromes. *Hum. Mol. Genet.* **20**, 880-893.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379-423.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498-2504.

Shaw, D. R. (2009). Searching the Mouse Genome Informatics (MGI) resources for information on mouse biology from genotype to phenotype. *Curr. Protoc. Bioinformatics* **1**, 1.7.

Slager, R. E., Newton, T. L., Vlangos, C. N., Finucane, B. and Elsea, S. H. (2003). Mutations in RAI1 associated with Smith-Magenis syndrome. *Nat. Genet.* **33**, 466-468.

Smith, C. L. and Eppig, J. T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **1**, 390-399.

Smith, C. L., Goldsmith, C. A. and Eppig, J. T. (2005). The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.* **6**, R7.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J. et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251-1255.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L. and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431-432.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., Lee, C. et al. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-853.

Tanaka, Y., Naruse, I., Maekawa, T., Masuya, H., Shiroishi, T. and Ishii, S. (1997). Abnormal skeletal patterning in embryos lacking a single Cbp allele: a partial similarity with Rubinstein-Taybi syndrome. *Proc. Natl. Acad. Sci. USA* **94**, 10215-10220.

Thorisson, G. A., Lancaster, O., Free, R. C., Hastings, R. K., Sarmah, P., Dash, D., Brahmachari, S. K. and Brookes, A. J. (2009). HGVbaseG2P: a central genetic association database. *Nucleic Acids Res.* **37**, D797-D802.

Venturin, M., Guarnieri, P., Natacci, F., Stabile, M., Tenconi, R., Clementi, M., Hernandez, C., Thompson, P., Upadhyaya, M., Larizza, L. et al. (2004). Mental retardation and cardiovascular malformations in NF1 microdeleted patients point to candidate genes in 17q11.2. *J. Med. Genet.* **41**, 35-41.

Verhoeven, W. M., Kleefstra, T. and Egger, J. I. (2010). Behavioral phenotype in the 9q subtelomeric deletion syndrome: a report about two adult patients. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **153B**, 536-541.

Vermeesch, J. R., Fiegler, H., de Leeuw, N., Szuhai, K., Schoumans, J., Ciccone, R., Speleman, F., Rauch, A., Clayton-Smith, J., Van Ravenswaaij, C. et al. (2007). Guidelines for molecular karyotyping in constitutional genetic diagnosis. *Eur. J. Hum. Genet.* **15**, 1105-1114.

Wang, D., Jao, L. E., Zheng, N., Dolan, K., Ivey, J., Zonies, S., Wu, X., Wu, K., Yang, H., Meng, Q. et al. (2007). Efficient genome-wide mutagenesis of zebrafish genes by retroviral insertions. *Proc. Natl. Acad. Sci. USA* **104**, 12428-12433.

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M. and Lewis, S. E. (2009). Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol.* **7**, e1000247.

Webber, C., Hehir-Kwa, J. Y., Nguyen, D. Q., de Vries, B. B., Veltman, J. A. and Ponting, C. P. (2009). Forging links between human mental retardation-associated CNVs and mouse gene knockout models. *PLoS Genet.* **5**, e1000531.

Zollino, M., Orteschi, D., Murdolo, M., Lattante, S., Battaglia, D., Stefanini, C., Mercuri, E., Chiurazzi, P., Neri, G. and Marangi, G. (2012). Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype. *Nat. Genet.* **44**, 636-638.